Cross-Domain Identifier Renaming Effects on CodeT5 Exact Match Accuracy in MBPP

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does cross-domain identifier renaming in training data impact CodeT5's exact match accuracy on the MBPP benchmark compared to single-domain fine-tuning. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: BEACON: Budget-Aware Entity Matching Across Domains (Extended Technical Report). Research question: How does cross-domain identifier renaming in training data impact CodeT5's exact match accuracy on the MBPP benchmark compared to single-domain fine-tuning?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.3/10.

3 Results

15 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 3.3/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
BEACON achieves a macro F1 score of 0.778 for a budget of 5k and 0.790 for a budget of 10k on the 50% CC WDC dataset.	×	0.05
SPEC, the second-best baseline, achieves a macro F1 score of 0.763 for a budget of 5k and 0.770 for a budget of 10k on t	×	0.04
The Cellphones category achieves an average F1 score of 0.980 across ten budgets despite having only 197 total samples.	×	0.02
The Computers category, the largest category, obtains a lower average F1 score of 0.787.	×	0.02
Most large domains exhibit more consistent F1 scores, typically in the 0.7–0.8 range.	×	0.05
The Automotive domain has 120 training samples with 57.4% positive, 32 validation samples with 23.7% positive, and 23 te	×	0.03
The Cameras domain has 1317 training samples with 57.7% positive, 204 validation samples with 17.7% positive, and 172 te	×	0.03
The Cell Phones domain has 151 training samples with 69.5% positive, 24 validation samples with 36.8% positive, and 48 t	×	0.03
The Clothing domain has 219 training samples with 58.2% positive, 28 validation samples with 21.2% positive, and 37 test	×	0.03
The Computers domain has 6473 training samples with 51.2% positive, 1174 validation samples with 14.6% positive, and 790	×	0.03

References

- http://arxiv.org/abs/2603.11391v2

- http://arxiv.org/abs/2212.02035v1
- http://arxiv.org/abs/2408.07888v2